古文書の翻刻支援システムの研究及び開発

古文書の翻刻支援システムの研究及び開発

STUDY ON DEVELOPING A SUPPORTING TOOL
FOR REPRODUCING CLASSIC HAND-WRITTEN DOCUMENTS

黒澤英博
Hidehiro KUROSAWA
指導教員　玉井哲雄

法政大学大学院理工学研究科システム理工学専攻修士課程

Because a variety of characters are used for one letter in old handwritten documents, it is hard to read those classic documents. To help people read such a document, it is effective to let a machine reprint it. The objective of this research is to build a program that recognizes handwritten old characters. There already exist studies to read ancient documents by character recognition, but the developed tools are widely available. Thus, there is a meaning to develop an application that performs character recognition with a smartphone. We propose a study of developing an application that photocopies classic documents with a smartphone camera and identifies characters from the image then translates them into the modern Japanese language.

***Key Words** : OCR, Ancient Document, Reprint*

１．はじめに

（１）背景

修士研究のテーマを決めるにあたり，身近な問題を解決できるプログラムを作成したいと考えた. そこで, 様々な人に話を伺ったところ，祖先が残した日記等の資料つまり古文書が読めないという話を聞いた．この問題を解決するプログラムを作成するにあたり解決方法を考えた.

古文書は一つの文字に対する字種が多く，筆による手書きなので読み取りづらい，加えて現代語と文法が違うため，自らが古文を学び読めるようになる方法では，学習に時間がかかってしまい，読むことができるまでが大変である．そこで，古文が読める人に頼んで訳してもらうことが考えられる．だがこの方法では自分と翻訳者との都合をあわせる必要があり，場合によっては自ら学習するより時間がかかる．その手間を省くために訳してもらう作業を機械にさせることが効率的だと考え，文字認識を用いることで解決するプログラムの作成を思いついた.

既に古文書を文字認識で読み解く研究はあるが，その文字認識を行うプログラムは販売・配布を行っているわけではないため普及しておらず，使用することができない[1],[2],[3]．そこで，研究者ではない一般の人々にもプログラムを利用できるように，現代社会において所持率の高いスマートフォンで文字認識を行うアプリケーションの開発を考えた.

すなわち研究内容はスマートフォンのカメラで古文書を撮影し，その画像から文字認識を行うことで現代語訳を得ることができるアプリケーションの研究及び開発である.

（２）目的

パターン認識には，人が発する声からその言葉を認識する音声認識，画像データの中から文字を認識する文字認識，画像に写る人の顔を判断することや指紋・虹彩を識別する個人認証などがある．パターン認識はこのように様々な形で使用され発達してきているが，その重要な一分野である文字認識について，その技術を把握し，またそれを応用したシステムを開発することが本研究の目的である.

２．手法

（１）文字認識方法

文字認識を行う際は，文字切り出しが重要となる．文字切り出しとは，画像中から文字列を認識し，次に一文字の範囲を認識することである．文字切り出しを行った後，切り出されたデータと辞書データとのマッチングを行い, マッチングの高い文字に対応する結果を出力する. 本研究で文字認識の対象とする古文書は筆で書かれていることもあり，くずし字だけでなく続け字がある．続け字とは，一文字毎に一定の間隔を空けて文字を書くのではなく，筆を離さずに次の文字を書くことである．古文書の文字認識は，続け字により文字切り出しが困難になっているため，完全自動で翻刻システムを作る際には，

続け字の問題を解決しなければならない．本研究では，完全自動のシステムではなく，あくまでも支援を行うシステムの開発を目指すため，対象の古文を一文字に切り出す作業はユーザに任せることとした．

本研究では，文字認識を行うために，既存の文字認識ソフトウェアを用いる．文字認識ソフトウェアには，オンライン文字認識とオフライン文字認識とがある．オンライン文字認識とは，タッチペンなどで入力された筆跡を追跡して，文字を動的に認識することである．一方，オフライン文字認識とは，紙などの上に書かれた文字の形をスキャニングし，文字を静的に認識することである．古文書は手書き文字であり，手書き文字に対応しているオンライン文字認識を用いることが適している面もあるが，本研究では既に書かれている文字を認識するため，オフライン文字認識を行うソフトウェアを用いる．文字認識には OCR（Optical Character Recognition,光学文字認識）技術が一般に用いられるが，古文に現れる特殊な文字を扱うため，認識方法や辞書データの変更を行う必要がある．そのため，市販の OCR ではなく，システムやデータの変更が自由に行えるオープンソースの OCR を用いる．オープンソースの OCR の中でも，既存の機能として日本語に対応しているものに，NHocr や Tesseract-OCR がある[4],[5],[6]．本研究では，この内の Tesseract-OCR を用いる．その理由は，NHocr では縦書きに対応しておらず，辞書データの変更もできないが，Tesseract-OCR ではそれらに対応しているためである．

古文に対応した辞書データを作成するにあたり，古文書翻刻支援システム開発プロジェクトから提供されている古文書文字データベース内の画像を用いて，Tesseract-OCR に対応させる[7]. 伏見屋善兵衛文書という文化～慶応（1804~1868）年間の資料から一文字ずつ切り出され，1,436 字種 142,663 文字からなるデータベースが提供されているが，これを用いる．

（２）古文翻刻方法

古文は文法が現代語と異なっているため OCR で認識した結果だけでは読むことは難しい．そのため OCR から出力された古文の文法を現代語の文法に変換するプログラムが必要になる．古文自動翻訳研究センターから古文から現代語に訳すフリーソフトウェアが提供されており，このソフトウェアに OCR で得られたテキストデータを入力することで，翻刻された現代語を得ることができ，認識対象の古文を読むことを可能とする[8].

（３）アプリケーション化

古文書が読めないという問題を解決するプログラムを普及するために，スマートフォンで古文書の文字認識を行うアプリケーションを開発する．アプリケーションを開発する環境として，Cordova を用いる．Cordova は Android・iOS のどちらでも動作するアプリケーションの開発を可能とするので，スマートフォンの使用 OS に関わらず普及させることができる．開発言語は HTML5 と JavaScript を用いる．

文字認識は，PC でさえ処理の時間がかかる．スマートフォンでは，更に時間がかかってしまうので，スマートフォンでプログラムを動作させるのではなく，サーバがスマートフォンから画像を受信し，プログラムを動作させ認識結果をスマートフォンに返すことで処理時間を短くする．

３．ツール開発と実験結果

認識対象の古文一文字の画像から文字認識を行い，一文字の認識結果を出力する．その後，繰り返し文字認識を行い，一文になった場合に翻刻機によって現代語へと翻刻する. この一連の動作を行うプログラムを作成した．

開発環境は，Eclipse pleiades，開発言語は Java である．

作成したプログラムの精度を確かめるために様々な方法で文字の認識を行う．スマートフォンにより送信される画像のサイズの違いによって認識の差が生じるため，Tesseract-OCR に適切な画像サイズを確認のための実験を行う．また，本研究で用いた手法である一文字ごとの認識の精度を確かめるため，Tesseract-OCR の既存の機能としての一度に一文を認識した結果と一文字ずつの認識結果との違いを比較する. 更に，辞書データをひらがな・カタカナと漢字に分け，認識対象にそれぞれ対応した字体の辞書データを用いた際に，認識精度がどのように向上するかを確かめる．

認識に使用する古文は，画像サイズと一文字ごとの精度差の実験では，茶席の禅語大辞典より宙宝宗宇(1759~1838 年)著「万里一条鉄」を，一文字ごとの精度さの実験では安政 2 年(1855 年)に平須賀村の百姓代・組頭・名手たちが書いた承諾書である「御請書之事」の最初の一文である「金七両也」，辞書データを分割した実験に嘉永 5 年(1852 年)に本田新田村名主の浜太郎により書かれた「送り一札之事」の「送り一札之事　一今般我等組下権次郎倅甚之助廿四才」という文と平安時代に書かれた竹取物語を慶長(1596~1615)年間で写本したものを再び写本した古活字十行本の冒頭のである「今は昔竹取の翁というものありけり」という文を使用する[9],[10].

図 1「万里一条鉄」

図 2「金七両也」

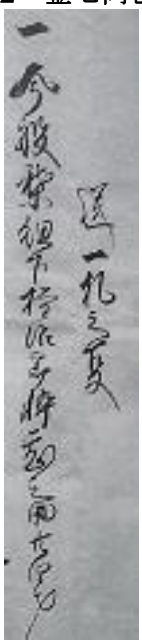

図 3「送り一札之事」

図 4「竹取物語」

表 1 「万里一条鉄」の画像サイズごとの認識結果

万里一条鉄	2100	2000	1900	1800	1700	1600	1500	1400	1300	1200	1100
万				薬		萬			普		萬
里								望		望	
一	返	婦	返	國	國	國	國	返	反	返	返
条	浚	幅	幅	幅	幅	幅	幅	幅	幅	幅	幅
鉄	魅	魅	魅	魅	魅	魅	魅	魅	魅	魅	魅

万里一条鉄	1000	900	800	700	600	500	400	300	200	100
万			嵜	変	萬	変	変	嵜	嵜	嵜
里		望	望	望	梁	望	望	望	墅	望
一	返	返	門	反	門	返	門	返	名	測
条	幅	幅	幅	繪	搗	幅	鈴	條	鈴	繪
鉄	魅	魅	魅	魅	魅	祝	魅	魅	魅	祝

表 2 一文字ずつと一文の認識結果

万里一条鉄	金七両也
一文字	嵜望返條魅　貪麦邊包
一文	氣寶幅魅　貪老彌 鳴賦戌

表 3 「送り一札之事」の認識結果

送り一札之事	送	一	札	之	事	一	今	般	我	等	組
	り										
認識結果	迚	艸	糺	奕	変	明	眷	紋	豫	讖	鍛
送り一札之事	下	権	次	郎	倅	甚	之	助	廿	四	才
認識結果	暮	検	併	壽	鰯	殺	忘	塲	冥	墅	煩

表 4 「竹取物語」の認識結果

竹取物語	い	ま	は	む	か	し	た	け	と	り	の
認識結果	わ	ひ	ろ	ね	わ	に	ゐ	び	テ	ろ	は
竹取物語	翁	と	い	う	も	の					
認識結果	輔	ざ	け	み	ぢ	わ					

４．考察・結論

文字認識に適切な画像サイズは，2100 ピクセルから 900 ピクセルまでの範囲では認識できない文字が存在し，結果が出力されず空白であった．800 ピクセルから 100 ピクセルまでの画像サイズは 5 文字すべての認識結果が出ている．加えて，正答している認識結果は 1600 ピクセル・1100 ピクセル・600 ピクセルの「万」と 300 ピクセルの「条」のみである．この結果より文字認識適している画像サイズは 600 ピクセルから 300 ピクセルの間であると確認できた．この実験以降で使用した画像は 300 ピクセルで文字認識をしている．

一文字と一文で認識した結果の精度向上の確認だが，どちらも正答した結果はないが，一文字ごとの認識は手法により字数は一致しているため，一文字ごとの認識の方が精度は向上している．

辞書データを分けた実行結果では，「送り一札之事」は「送」と「迸」・「札」と「糺」，「権」と「桧」など部首が同じものもあったが正答した認識結果はなかった．「竹取物語」は辞書データを分割有り無しどちらとも正答はなかった．辞書データを分割しない場合だとひらがなに対しても漢字と認識された．

結論として，文字認識の精度はかなり低く，翻刻がうまくいかなかった．辞書データを分割しない際の「竹取物語」の結果であるが，ひらがなに対して漢字の結果が出ている．これは，辞書データに使用した古文書文字データベースに収録されているひらがな・カタカナの数が約 9,000 件と少なく，漢字は約 130,000 件と収録数の違いによると思われる．このような学習の差をなくすため，辞書データを分割して認識を行ったが結果は芳しくなかった．この理由として思い当たるのは，ひとつはひらがな・カタカナ・漢字だけでなく，１つの字種に対しての収録数が違うことである．もうひとつは，認識対象の「竹取物語」の時代が学習した文字の時代と違いであると考えられる．

既存の言語を用いた Tesseract-OCR であれば，文字認識の際に文法を学習でき，そこから予測して精度を向上することが可能である．だが本研究では，古文書の難点である文字切り出しをユーザの手作業で行い，一文字ずつ認識したことで文法からの精度向上が利用できなかった．

今後の展望として，今回辞書データと同じ時代の古文の数を揃えられなかったため，認識対象を増やし，どのような文字の字体が認識しやすいかを理解することで学習データを適したものに変更し，精度の向上を目指したい．加えて， Tesseract-OCR の機能を十分に使用できるよう，文字切り出しを行えるシステムを考えなければならない．

５．参考文献

1) ベロフ　アレクサンドル，立田ルミ:漢字 OCR システムでの認識率の向上方法と考察 -携帯機器上の問題点と解決方法-，情報学研究，Feb.2012

2) 寺沢憲吾，長崎健，川島稔夫:固有空間法と DTW による古文書ワードスポッティング，電子情報通信学会，pp1829-1839，2006

3) 朱碧蘭，中川正樹:オンライン手書き文字認識の最新動向，電子情報通信学会，Vol.95 No.4，pp335-340，2012

4) tesseract-ocr · GitHub，https://github.com/tesseract-ocr

5) NHocr: 日本語文字認識プログラム，https://ja.osdn.net/projects/nhocr/

6) 小沼元輝，朱碧蘭，山田奨治，柴山守，中川正樹:古文書解読を支援するくずし字辞典のための文字認識の開発，情報考古学，Vol.13 No.1，pp.22-33，2007

7) HCR Project，http://ys.nichibun.ac.jp/~shoji/hcr/

8) 古文自動翻訳研究センター，http://honnyaku.okunohosomichi.net/

9) 有馬頼底：茶席の禅語大辞典，淡交社，p951，2002

10) 片桐洋一：竹取翁物語：古活字十行本，和泉書院，1986

11) 山田奨治，柴山守:n-gram と OCR による定型表現がある古文書の文字の推定，情報処理学会研究報告書，2003

12) 渡邊一義，渡辺悟，鈴木徹也:古文書文字認識における N-gram の出現頻度を利用した認識候補削減法,情報処理学会第 77 回全国大会，2015

13) 尾崎浩司，柴山守，荒木義彦：古文書画像のレイアウト認識と表題抽出，人文科学とコンピュータ研究会，p 47-54，2000

14) 高田智和，矢田勉，斎藤達哉：変体仮名のこれまでとこれから，情報管理 9 月号 vol.58，p438-446，2015

15) 寺沢憲吾，川嶋稔夫：文書画像からの全文検索のオンラインサービス，人文科学とコンピュータシンポジウム，p329-334，2011

16) 小高和巳，若原徹，増田功：筆順に依存しないオンライン手書き文字認識アルゴリズム，電子通信学会論文誌，vol.J65-D，p679-686，1982

17) 大森健児:続け字と崩し字に対応したヒューリスティックなストローク合わせ法によるオンライン手書き漢字認識，情報処理学会論文誌，vol.31 No.5，p710-720，1990

18) Tess4J - JNA wrapper for Tesseract，http://tess4j.sourceforge.net/

19) Apache Cordova，https://cordova.apache.org/